CLAIMS:

1. An apparatus for identifying one or more portions of data in a database for comparison with a query input by a user, the query and the portions of data each comprising a sequence of sub-word units, the apparatus comprising:

a memory for storing data defining a plurality of sub-word unit classes, each class comprising sub-word units that are confusable with other sub-word units in the same class;

a memory for storing an index having a plurality of entries, each of which comprises:

(i) an identifier for identifying the entry;

(ii) a key associated with the entry and which is related to the identifier for the entry in a predetermined manner; and

(iii) a number of pointers which point to portions of data in the database which correspond to the key for the entry;

wherein each key comprises a sequence of sub-word unit classifications which is derived from a corresponding sequence of sub-word units appearing in the database by classifying each of the sub-word units in the sequence into one of the plurality of sub-word unit

classes;

means for classifying each of the sub-word units in the input query into one of the plurality of sub-word unit classes and for defining one or more sub-sequences of query sub-word unit classifications;

means for determining a corresponding identifier for an entry in said index for each of said one or more sub-sequences of query sub-word unit classifications;

means for comparing the key associated with each of the determined identifiers with the corresponding sub-sequence of query sub-word unit classifications; and

means for retrieving one or more pointers from said index in dependence upon the output of said comparing means, which one or more pointers identify said one or more portions of data in the database for comparison with the input query.

2. An apparatus according to claim 1, wherein said sub-word units are phonemes or phoneme-like units.

3. An apparatus according to claim 1, wherein at least ten sub-word unit classes are defined in advance.

4. An apparatus according to claim 1, wherein each key is related to the corresponding identifier by a

predetermined mathematical function.

5. An apparatus according to claim 4, wherein each key is related to the corresponding identifier by the following equation:

$$\left[\prod_{i=1}^{W}\left[C[i]K_c\right]\right] Mod\ S$$

where K_c is the number of sub-word unit classes, S is the number of entries in the index, C[i] is the number of the sub-word class to which the i^{th} sub-word unit in the sequence of sub-word units corresponding to the key belongs and W is the number of sub-word unit classifications in each key.

6. An apparatus according to claim 1, wherein said determining means is operable to identify a new identifier for another entry in said index for a subsequence of query sub-word unit classifications if said comparing means determines that the key for the identifier is not the same as the subsequence of query sub-word unit classifications.

7. An apparatus according to claim 6, wherein said determining means is operable to determine a new

identifier using the following equation:

$$IDX = [IDX + V] \, Mod \, S$$

where (IDX) is the identifier, S is the number of entries in the index and V is a predetermined number.

8. An apparatus according to claim 1, wherein the key for one or more of said entries is a null key indicating that there are no pointers stored in the index for that entry.

9. An apparatus according to claim 4, wherein said determining means is operable to determine a corresponding identifier for each subsequence of query sub-word unit classifications using said predetermined mathematical function.

10. An apparatus according to claim 1, wherein said input query is a typed query and wherein the apparatus further comprises means for converting the typed query into said sequence of sub-word units.

11. An apparatus according to claim 1, wherein said input query is a spoken query and wherein the apparatus

further comprises a speech recognition system for processing the spoken query and for outputting said sequence of subword units.

12. An apparatus for searching a database in response to a query input by a user, the database comprising a plurality of sequences of sub-word units and the query comprising at least one sequence of sub-word units, the apparatus comprising:

an apparatus according to any of claims 1 to 11 for identifying one or more portions of data in the database for comparison with the input query; and

means for comparing the one or more sequences of query sub-word units with the identified one or more portions of data in said database.

13. An apparatus according to claim 12, wherein said means for comparing said input query with said portions of data in the database uses a dynamic programming comparison technique.

14. An apparatus according to claim 12, further comprising means for retrieving one or more data files in dependence upon the results of said comparing means.

15. An apparatus for identifying one or more portions of data in a database for comparison with a query input by a user, the query and the portions of data each comprising a sequence of features, the apparatus comprising:

a memory for storing data defining a plurality of feature classes, each class comprising features that are confusable with other features in the same class;

a memory for storing an index having a plurality of entries, each of which comprises:

(i) an identifier for identifying the entry;

(ii) a key associated with the entry and which is related to the identifier for the entry in a predetermined manner; and

(iii) a number of pointers which point to portions of data in the database which correspond to the key for the entry;

wherein each key comprises a sequence of feature classifications which is derived from a corresponding sequence of features appearing in the database by classifying each of the features in the sequence into one of the plurality of feature classes;

means for classifying each of the features in the input query into one of the plurality of feature classes and for defining one or more sub-sequences of query

feature classifications;

means for determining a corresponding identifier for an entry in said index for each of said one or more sub-sequences of query feature classifications;

means for comparing the key associated with each of the determined identifiers with the corresponding sub-sequence of query feature classifications; and

means for retrieving one or more pointers from said index in dependence upon the output of said comparing means, which one or more pointers identify said one or more portions of data in the database for comparison with the input query.

16. Data defining an index for use in searching a database, the data comprising:

data defining a respective identifier for each of a plurality of entries in the index;

data defining a respective key for each of the plurality of entries, which keys are related to the corresponding identifiers in a predetermined manner; and

data defining a respective one or more pointers for a plurality of the entries, which pointers point to locations within the database corresponding to the key for the entry;

wherein each key comprises a sequence of sub-word

unit classifications which is derived from a corresponding sequence of sub-word units appearing in the database by classifying each of the sub-word units in the sequence into one of a plurality of sub-word unit classes, the sub-word unit classes being defined in advance and each comprising sub-word units that are confusable with other sub-word units in the same class.

17. A method of identifying one or more portions of data in a database for comparison with a query input by a user, the query and the portions of data each comprising a sequence of sub-word units, the method comprising the steps of:

storing data defining a plurality of sub-word unit classes, each class comprising sub-word units that are confusable with other sub-word units in the same class;

storing an index having a plurality of entries, each of which comprises:

(i) an identifier for identifying the entry;

(ii) a key associated with the entry and which is related to the identifier for the entry in a predetermined manner; and

(iii) a number of pointers which point to portions of data in the database which correspond to the key for the entry;

wherein each key comprises a sequence of sub-word unit classifications which is derived from a corresponding sequence of sub-word units appearing in the database by classifying each of the sub-word units in the sequence into one of the plurality of sub-word unit classes;

classifying each of the sub-word units in the input query into one of the plurality of sub-word unit classes and for defining one or more sub-sequences of query sub-word unit classifications;

determining a corresponding identifier for an entry in said index for each of said one or more sub-sequences of query sub-word unit classifications;

comparing the key associated with each of the determined identifiers with the corresponding sub-sequence of query sub-word unit classifications; and

retrieving one or more pointers from said index in dependence upon the output of said comparing step, which one or more pointers identify said one or more portions of data in the database for comparison with the input query.

18. A method according to claim 17, wherein said sub-word units are phonemes or phoneme-like units.

19. A method according to claim 17, wherein at least ten sub-word unit classes are defined in advance.

20. A method according to claim 17, wherein each key is related to the corresponding identifier by a predetermined mathematical function.

21. A method according to claim 20, wherein each key is related to the corresponding identifier by the following equation:

$$\left[\prod_{i=1}^{W}\left[C[i]K_c\right]\right] Mod\ S$$

where K_c is the number of sub-word unit classes, S is the number of entries in the index, C[i] is the number of the sub-word class to which the i^{th} sub-word unit in the sequence of sub-word units corresponding to the key belongs and W is the number of sub-word unit classifications in each key.

22. A method according to claim 17, wherein said determining step identifies a new identifier for another entry in said index for a subsequence of query sub-word unit classifications if said comparing step determines that the key for the identifier is not the same as the

subsequence of query sub-word unit classifications.

23. A method according to claim 22, wherein said determining step determines a new identifier using the following equation:

$$IDX = [IDX + V] \, Mod \, S$$

where IDX is the identifier, S is the number of entries in the index and V is a predetermined number.

24. A method according to claim 17, wherein the key for one or more of said entries is a null key indicating that there are no pointers stored in the index for that entry.

25. A method according to claim 23, wherein said determining step determines a corresponding identifier for each subsequence of query sub-word unit classifications using said predetermined mathematical function.

26. A method according to claim 17, wherein said input query is a typed query and wherein the method further comprises the step of converting the typed query into said sequence of sub-word units.

27. A method according to claim 17, wherein said input query is a spoken query and wherein the method further comprises the step of using a speech recognition system to process the spoken query to generate said sequence of subword units.

28. A method of searching a database in response to a query input by a user, the database comprising a plurality of sequences of sub-word units and the query comprising at least one sequence of sub-word units, the method comprising:

the method steps of claim 17 for identifying one or more portions of data in the database for comparison with the input query; and the step of

comparing the one or more sequences of query sub-word units with the identified one or more portions of data in said database.

29. A method according to claim 28, wherein said comparing step uses a dynamic programming comparison technique to compare the input query with said portions of data.

30. A method according to claim 28, further comprising the step of retrieving one or more data files in

dependence upon the results of said comparing step.

31. A storage medium storing processor implementable instructions for controlling a processor to implement the method of claim 17 or storing the data of claim 16.

32. Processor implementable instructions for controlling a processor to implement the method of claim 17.

33. An apparatus for identifying one or more portions of data in a database for comparison with a query input by a user, the query and the portions of data each comprising a sequence of sub-word units, the apparatus being characterised by an index having a plurality of entries, each of which includes a key comprising a sequence of sub-word unit classifications, which key is derived from a corresponding sequence of sub-word units appearing in the database by classifying each of the sub-word units in the sequence into one of a plurality of sub-word unit classes, each class comprising sub-word units that are confusable with other sub-word units in the same class.